

03-27-00

A

Please type a plus sign (+) inside this box → ☐

Approved for use through 09/30/2000 OMB 0651-0032  
 Patent and Trademark Office U.S. DEPARTMENT OF COMMERCE  
 Under the Paperwork Reduction Act of 1995 no persons are required to respond to a collection of information unless it displays a valid OMB control number

03/24/00  
 3715 U.S. PTO

# UTILITY PATENT APPLICATION TRANSMITTAL

(Only for new nonprovisional applications under 37 C.F.R. § 1.53(b))

Attorney Docket No. VGEN.P-056-US  
 First Inventor or Application Identifier IZMAILOV, ET AL  
 Title Method for Alignment of DNA Sequences ...  
 Express Mail Label No. EL362858548US

## APPLICATION ELEMENTS

See MPEP chapter 600 concerning utility patent application contents.

1. ☒ \* Fee Transmittal Form (e.g., PTO/SB/17)  
 (Submit an original and a duplicate for fee processing)
2. ☒ Specification [Total Pages 22]  
 (preferred arrangement set forth below)  
 - Descriptive title of the Invention  
 - Cross References to Related Applications  
 - Statement Regarding Fed sponsored R & D  
 - Reference to Microfiche Appendix  
 - Background of the Invention  
 - Brief Summary of the Invention  
 - Brief Description of the Drawings (if filed)  
 - Detailed Description  
 - Claim(s)  
 - Abstract of the Disclosure
3. ☒ Drawing(s) (35 U.S.C. 113) [Total Sheets 6]
4. Oath or Declaration [Total Pages]  
 a. ☐ Newly executed (original or copy)  
 b. ☐ Copy from a prior application (37 C.F.R. § 1.63(d))  
 (for continuation/divisional with Box 16 completed)  
 i. ☐ DELETION OF INVENTOR(S)  
 Signed statement attached deleting  
 inventor(s) named in the prior application,  
 see 37 C.F.R. §§ 1.63(d)(2) and 1.33(b).

## ADDRESS TO:

Assistant Commissioner for Patents  
 Box Patent Application  
 Washington, DC 20231

5. ☐ Microfiche Computer Program (Appendix)
6. Nucleotide and/or Amino Acid Sequence Submission  
 (if applicable, all necessary)  
 a. ☐ Computer Readable Copy  
 b. ☐ Paper Copy (identical to computer copy)  
 c. ☐ Statement verifying identity of above copies

## ACCOMPANYING APPLICATION PARTS

7. ☐ Assignment Papers (cover sheet & document(s))
8. ☐ 37 C.F.R. § 3.73(b) Statement of Power of Attorney  
 (when there is an assignee)
9. ☐ English Translation Document (if applicable)
10. ☐ Information Disclosure Statement (IDS)/PTO-1449 ☐ Copies of IDS Citations
11. ☐ Preliminary Amendment
12. ☒ Return Receipt Postcard (MPEP 503)  
 (Should be specifically itemized)
13. ☐ \* Small Entity Statement(s) ☐ Statement filed in prior application,  
 (PTO/SB/09-12) Status still proper and desired
14. ☐ Certified Copy of Priority Document(s)  
 (if foreign priority is claimed)
15. ☐ Other: \_\_\_\_\_

\* NOTE FOR ITEMS 1 & 13: IN ORDER TO BE ENTITLED TO PAY SMALL ENTITY  
 FEES, A SMALL ENTITY STATEMENT IS REQUIRED (37 C.F.R. § 1.27), EXCEPT  
 IF ONE FILED IN A PRIOR APPLICATION IS RELIED UPON (37 C.F.R. § 1.28).

## 16. If a CONTINUING APPLICATION, check appropriate box, and supply the requisite information below and in a preliminary amendment:

☐ Continuation ☐ Divisional ☐ Continuation-in-part (CIP) of prior application No. \_\_\_\_\_ / \_\_\_\_\_  
 Prior application information. Examiner \_\_\_\_\_ Group / Art Unit: \_\_\_\_\_

For CONTINUATION or DIVISIONAL APPS only: The entire disclosure of the prior application, from which an oath or declaration is supplied under Box 4b, is considered a part of the disclosure of the accompanying continuation or divisional application and is hereby incorporated by reference. The incorporation can only be relied upon when a portion has been inadvertently omitted from the submitted application parts.

## 17. CORRESPONDENCE ADDRESS

☒ Customer Number or Bar Code Labelor ☐ Correspondence address below

(Insert Customer No. or Attach bar code label here)

Name	PATENT TRADEMARK OFFICE			
Address				
City	State	Zip Code		
Country	Telephone	970-668-2050	Fax	970-668-2082

Name (Print/Type)	Marina T. Larson, Ph.D.	Registration No. (Attorney/Agent)	32,038
Signature	Marina T. Larson	Date	3/24/00

Burden Hour Statement. This form is estimated to take 0.2 hours to complete. Time will vary depending upon the needs of the individual case. Any comments on the amount of time you are required to complete this form should be sent to the Chief Information Officer, Patent and Trademark Office, Washington, DC 20231. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO Assistant Commissioner for Patents, Box Patent Application, Washington, DC 20231.

EL362858548US

Method for Alignment of DNA Sequences with  
Enhanced Accuracy and Read Length

Background of the Invention

This application relates to DNA sequencing technology and in particular to a method for alignment of DNA sequences which provides enhanced accuracy and read-length.

5 DNA sequencing is generally performed today using one of two methodologies: the chemical degradation method or the chain termination method. Of these, the chain termination method originally described by Sanger et al., *Proc. Natl. Acad. Sci. USA* 74: 5463-5467 (1977) or variations thereof have been adopted in many cases for development of automated sequencing instruments and protocols.

In the chain termination sequencing method, fragments are generated using chain termination reagents in a template-dependant polymerization reaction. The lengths of the fragments indicate the positions of one species of base in a target polynucleotide. If fragment sets are generated for each of the four species of bases (A, C, G and T), analysis of the fragment sizes permits the explicit determination of the sequence of the target polynucleotide. While the translation of this conceptual methodology into practice is effective for determination of sequences, the application in automated systems has faced numerous challenges. These include the fact that the band shape produced following electrophoresis of real fragments is not consistent from one band to the next and may not be perfectly straight (smiling may occur); variations which can occur in peak spacing from one lane of a gel to the next; variations in peak spacing which can occur as the length of the run increases; and decreases in resolution which occur as the length of the run increases. Furthermore, since much of the cost associated with DNA sequencing is in the set-up time involved, for clinical and diagnostic applications the larger the length of DNA which can be sequenced with accuracy, the smaller the per patient cost can be. These considerations have led to a variety of proposals for improving the chemistry used in

sequencing, or for improving the manner in which data representing the detected sequencing fragment is processed. The present invention relates to the second type of improvement.

In order to obtain meaningful sequence information from raw data obtained by electrophoresis of labeled sequencing fragments, one of the most important factors is the alignment of the data traces representing each species of base. In non-automated systems, this is frequently done by eye-ball, and the eye of a skilled technician is in fact a remarkable tool for this purpose. Commonly assigned US Patent No. 5,916,747, which is incorporated herein by reference, discloses a method for aligning data traces from four channels of an automated electrophoresis detection apparatus in which each channel detects the products of one of four chain-termination DNA sequencing reactions such that the four channels together provide information concerning the sequence of all four bases within a nucleic acid polymer being analyzed. The method places the four data traces in a trial alignment, and then determines coefficients of shift and stretch for selected data points within each normalized data trace to optimize a cost function which reflects the extent of overlap of peaks in the combined normalized data traces to which the coefficients have been applied. Warp functions are then generated for the normalized data traces from the coefficients of shift and stretch determined for the selected data points, and applied to the respective data trace to produce four warped data traces which are assembled to form an aligned data set. This data set is then used for base-calling to complete the sequence determination process.

The procedure of the '747 patent is generally suited for the determination of sequences where explicit data for the positions of all four bases are obtained. On the other hand, it is not always necessary to determine the positions of all of four species of bases in order to obtain diagnostic information from a given polynucleotide. (See, commonly assigned US Patent No. 5,834,189, which is incorporated herein by reference). Commonly assigned US Patent No. 5,853,979, which is incorporated herein by reference discloses a method for the interpretation of experimental fragment patterns for polynucleotides having putatively known sequences. In this method, at least one raw fragment pattern representing the positions of a selected nucleotide base

as a function of migration time or distance is obtained for the experimental sample. The fragment pattern is evaluated to determine one or more "normalization coefficients." These normalization coefficients reflect the displacement, stretching or shrinking, and rate of stretching or shrinking of the clean fragment, or segments thereof, which are necessary to obtain a suitably high degree of correlation between the clean fragment pattern and a standard fragment pattern which represents the positions of the selected nucleic acid base within a standard polymer actually having the known sequence as a function of migration time or distance. The normalization coefficients are then applied to the fragment pattern to produce a normalized fragment pattern which is used for base-calling in a conventional manner. As indicated, however, this technique requires prior knowledge of the expected fragment pattern for the polynucleotide being analyzed.

Notwithstanding such techniques, there remains room for improvement in the manner in which automated analysis of sequencing fragment patterns are carried out. In particular, there remains a need for systems which allow enhanced read-length, i.e., the analysis of a greater number of bases in a single lane of a gel, without loss of accuracy or substantial increase in analysis time. It is an object of the present invention to provide a method which answers this need.

#### Summary of the Invention

The present invention provides a method for aligning sequence data traces. In accordance with the invention, an experimental data trace representing the positions of a first species of base within a target polynucleotide and a reference data trace representing the positions of a second species of base (which may be the same as or different from the first species) within a reference polynucleotide are obtained by separating appropriate sequencing fragments generated from the target and reference polynucleotides in a common lane of an electrophoresis gel. For each reference data trace, a plurality of peaks corresponding to fragments having a size in the range of 40 to 1200 bases are selected. A base number is assigned

to each of the selected peaks in the reference data trace, and a numerical “peak file” is created with information about the peak number and migration time (or distance). This peak file is analyzed to determine a set of polynomial coefficients which will allow substantial linearization of a plot of peak number versus separation between adjacent peaks and alignment of the traces with respect to each other. These coefficients are used to create a corrected time scale identifying where peaks should be located on a given experimental gel. This corrected time scale is used to guide the sampling of the experimental data, and for assignment of peaks within the data.

#### Brief Description of the Drawings

Fig. 1 shows a plot of peak spacing versus peak number for unaligned data, and data aligned with third and fifth order polynomials;

Fig. 2 shows a plot of peak spacing versus peak number for data aligned with third, fourth and fifth order polynomials;

Figs. 3A and B show plots of the difference, for each lane, between the run time of a base (322nd nt) and its average value for all 16 lanes of a gel. Fig. 3A corresponds to the run time difference in the raw data; Fig. 3B is the run time difference after alignment;

Fig. 4 shows the relationship between accuracy and read length for a first set of experimental data which was well-aligned on the gel;

Fig. 5 shows the relationship between accuracy and read length for a first set of experimental data which was poorly-aligned on the gel; and

Fig. 6 shows a system in accordance with the invention.

#### Detailed Description of the Invention

The present invention provides a method for linearization and alignment sequence data traces. As used herein, the term “linearize” refers to establishing equal spacing in a time domain between adjacent peaks within the overall sequence in an experimental data trace. The term “align” refers to establishing the correct positions within the overall sequence for the peak

in an experimental data trace. When a data trace is obtained for each of the four bases, the alignment process results in an explicit determination of the position of each and every base. However, since in some instances it is not necessary to perform all four sequencing reactions and analyze the results to obtain useful diagnostic data, "alignment" can be performed on a single  
5 trace, representing the positions of a single species of nucleotide base within a target polynucleotide. In this case, the single trace after linearization is "aligned" with a standard time scale, to show the base numbers associated with peaks within the linearized trace. Alignment can also be performed on data sets of two or more traces representing the positions of two or more species of nucleotide bases within the target polynucleotide.

10 The process of linearization and alignment is essentially one of assigning a correct numerical position to each of the bases. An important aspect of the linearization and alignment process is compensation for variation in peak spacing which occurs over time even within a single lane of an electrophoresis gel. The present invention performs this compensation by co-electrophoresing a reference sequence with the experimental sequence and utilizing the resulting  
15 reference data trace to define the correct peak spacing.

20 The specification and claims of this application use the term "DNA sequencing fragments" to describe the mixture of polynucleotides which results when chain extension polymerization is performed in the presence of a chain-terminating base analog, such as a dideoxynucleotide triphosphate. The term "DNA sequencing fragments" only requires the presence in the mixtures of fragments the lengths of which are indicative of the positions of one type of base within the polynucleotide being analyzed.

25 In the simplest embodiment of the invention, experimental and reference data traces obtained from a single lane of an electrophoresis gel are evaluated. The experimental polynucleotide may be, for example, the A-sequencing fragments generated from a target polynucleotide of interest. The reference sample is, for example, the T-sequencing fragments generated from a reference polynucleotide of known sequence. Preferably, the reference

polynucleotide is of similar total length to the experimental polynucleotide so that the reference data extends over the entire length of the experimental sequence information.

Because the reference polynucleotide has a known sequence, it is possible to immediately create a peak table having two columns: actual retention time and peak number.

Thus, for example, if the sequence were 9 bases long, and had the sequence ACATTACGA, then the data trace derived from the A-sequencing fragment would have four peaks appearing at times  $T_1$ ,  $T_2$ ,  $T_3$  and  $T_4$ , respectively. The peak table would therefore appear as follows:

$T_1$	1
$T_2$	3
$T_3$	6
$T_4$	9

If the spacing of the peaks in the gel over this region were exactly the same, then a plot of  $T$  versus peak number would produce a straight line, and a plot of the spacing (the difference between each adjacent peak) versus peak number would produce a straight, horizontal line.

Because experimental data does not meet this ideal, however, the result is in fact far different. Thus, as shown in Fig. 1, the experimental spacing between adjacent peaks as a function of base number may follow a complex curve, at first increasing through a maximum, and then decreasing again.

In accordance with the present invention, a curve fitting procedure is applied to the raw reference data trace in which the data is fit to a polynomial, generally a third or higher-order polynomial. Although this fitting process is generally performed in actual practice using a computer program and any of various known curve fitting programs, the procedure employed can be understood from the discussion below. In the unaligned data, one is essentially plotting the function

$$\Delta T = mP + c$$

where  $\Delta T$  is the spacing between adjacent peaks (in units of time),  $m$  is the slope of the line,  $c$  is a constant which is characteristic of the gel and which reflects the characteristic peak spacing,

and P is peak number. In ideal data, the slope m is 0, such that there is no actual relationship between  $\Delta T$  and P, and  $\Delta T$  is simply a constant. As one can see in Fig. 1, the experimental data are far from being a straight line. In this case, the experimental curve can be approximated by a polynomial. The empirical curve  $\Delta T$  is fit to a polynomial function by a least squares method:

$$\Delta T = a_{i0} + a_{i1}P + a_{i2}P^2 + \dots + a_{ik}P^k.$$

The degree (k) of the polynomial is an input parameter of the fitting program. The procedure generates a set of coefficient  $\{a_{ik}\}$  for each gel lane (i). A curve fitting program identifies the coefficients,  $a_{ik}$ , (which may be positive or negative) and the constant  $a_{i0}$  which bring the resulting plot of the reference data closest to a straight line. Based on the set of polynomial coefficients,  $\{a_{ik}\}$ , a corrected time scale is defined for each peak in gel lane #i, according to the formula

$$T_{ip} = C_i [a_{i0} + a_{i1}t_{ip}P + a_{i2}t_{ip}P^2 + \dots + a_{ik}t_{ip}P^k],$$

where  $T_{ip}$  is the corrected time value for the reference peak of length p,  $t_{ip}$  is the experimentally measured run time of this peak, and  $C_i$  is a scaling factor. This transformation causes the spacing between consecutive peaks in the corrected time domain ( $dT_{ip}/dp$ ) to remain constant over the course of the run. The transformation (linearization) is performed for both the reference peaks and the sample peaks in each gel lane (i).

Each gel lane has a different scaling factor,  $C_i$ . For any particular gel, the set of values  $\{C_i\}$  is chosen to equalize the spacing between consecutive peaks in the corrected time domain, ( $dT_{ip}/dp$ ), across all lanes of the gel. A gel lane is uniformly compressed by setting  $C_i < 1$ , and it is uniformly stretched by setting  $C_i > 1$ . A set of coefficients  $\{C_i\}$  is therefore defined, such that all lanes of the gel have the same total run time in the correct time domain. In the dimension of real time, the data points are evenly spaced. However, in the dimension of “corrected time”, the corresponding time intervals are not of equal lengths. Therefore the experimental data sets are “resampled” into equally-spaced values (in the corrected time domain)



by quadratic interpolation. Resampling of the data set for each gel lane is done separately, because the corrected time scale may be different for each lane. The procedure described above is a global alignment, which precedes any subsequent local alignment by the base calling software. This global alignment procedure is general, and should be compatible with all types of local alignment algorithms.

The basic methodology described above for alignment of a single data trace can also be applied in other embodiments. For example, data can be obtained for all four bases (A,C, G and T) in four lanes, to obtain explicit position information for the complete sequence of a target polynucleotide. In this case, a set of reference sequencing fragments is desirably run in each of the four lanes. Further, in multi-lane gels, it is desirable to run a set of reference sequencing fragments in each lane, regardless of the nature of the experimental samples. If a sequencing apparatus is used that is capable of distinguishing between more than two labels, multiple experimental sets of sequencing fragments may be run in one lane along with a set of reference sequencing fragments. In each case where more than one reference data trace is obtained from a gel, the spacings of all the reference data traces can be combined to produce a single set of coefficients and single characteristic spacing which is applied to all of the experimental data traces from the gel.

Several features are common to all of the various embodiments discussed above. Each set of the experimental sequencing fragments and the reference sequencing fragments are labeled with a distinguishable labels, i.e, the labels on the experimental fragments and reference fragments are different from one another when they are present in the same lane of the gel. The nature of the labels is a matter of choice and compatibility with the detection system employed. Suitable labels include radiolabels, chromophores, chromogenic labels and fluorogenic labels. Preferred labels, however, are fluorescent labels compatible with automated multi-dye sequencers. Specific examples of suitable fluorescent labels include cyanine dyes such as Cy5.0 and Cy5.5 (See US Patents Nos. 4,981,977 and 5,268,486) and energy transfer dyes (U.S. Patent No. 5,800,996) and rhodamine dyes (U.S. patents Nos. 5,366,860 and 4,855,225).

There is no required relationship between the target polynucleotide and the reference polynucleotide, and it is not mandatory that the same set of reference sequencing fragments be used in all of the lanes of a gel. This is the case because the alignment depends on the measured position of the known bases of the reference trace, but not on the identity of the bases. However, the reference polynucleotide should be selected to provide enough peaks (or bands) to facilitate the use of a desirable degree of polynomial for fitting the experimental data. For example, if one wishes to use a 5th-degree polynomial, the reference polynucleotide must provide at least 6 peaks.

Furthermore, while it is necessary to know the sequence of the reference polynucleotide for the creation of the initial peak table, it is not necessary to have any *a priori* knowledge of the sequence of the target polynucleotide. Thus, while the present invention is particularly applicable to diagnostic applications where the putative sequence of the target polynucleotide is known, it is not limited to such applications.

A further factor which can be adjusted by the user is the number of peaks within the reference data trace that are used in determining the polynomial coefficients and characteristic spacing. While all of the peaks can be considered, this increases the processing time and burden. As a practical matter, a much smaller number of peaks can be utilized and still provide good alignment of the experimental data traces. For example, for alignment of sequencing fragments spanning 40 to 1,200 bases, from 3 to 40 peaks in the reference data trace are suitably selected. The selected peaks are preferably distributed fairly evenly throughout the reference data trace, although precisely equal distribution is not required.

Fig. 6 shows a schematic representation of an apparatus in accordance with the present invention for evaluating the sequence of a target polynucleotide. The apparatus as shown comprises a processor housing 10 which has an input 11 for receiving information about one or more experimental DNA sequencing data traces derived from the separation of experimental DNA sequencing fragments reflecting the position of at least one base in the target polynucleotide and one or more reference DNA sequencing data traces derived from the

separation of reference DNA sequencing fragments reflecting the position of at least one base in a reference polynucleotide of known sequence. For example, input 11 may be in the form of a wire for transmitting sequence-related data from a sequencer. Data could also be transmitted via a wireless link, or communicated to the apparatus through disk drive 13.

5                Within the housing 10 is a data processing apparatus 14 which include one or several processors. The processors or processors are operatively programmed

              (a)        to evaluate the reference DNA sequencing data traces to determine a corrected time scale indicative of migration times at which peaks should occur;

              (b)        to sample the experimental DNA sequencing data traces at time points  
10                determined by the corrected time scale; and

              (c)        to assign a base number to each peak found in the experimental DNA sequencing data traces based upon the corrected time scale, thereby obtaining information about the sequence of the target polynucleotide. The assigned base numbers may be further processed to provide an output indicative of information about the sequence of the target polynucleotide and this information is communicated to the user via an output device. Exemplary output  
15                devices are a display 15 or printer 16. The information may also be communicated by saving it to the disk drive 13 (which can function as either an input or an output device) or through a telecommunication connection (such as a modem or internet connection).

              In an embodiment of the invention, the processor programmed to evaluate the  
20                reference DNA sequence data traces is programmed to perform the steps of:

              (i)        identifying a plurality of peaks in the reference DNA sequencing data traces, and creating a data table containing the number of each peak based on the known sequence of the polynucleotide, and the position of each peak in the reference DNA sequencing data trace;

25                (ii)        identifying a set of coefficients for a polynomial effective to substantially linearize a plot of peak number versus separation between adjacent peaks; and

(iii) creating from the coefficients and the polynomial a corrected time scale which reflects the positions at which a peak should occur at any given point in a sequencing data trace.

The invention will now be further described and illustrated with reference to the following, non-limiting examples.

#### Example 1

Lanes 1, 5, 9 and 13 of a standard 16 lane MICROCEL™ electrophoresis gel (Visible Genetics Inc.) were loaded with a mixture of the A-terminated sequencing fragments from M13 labeled with CY5.0 fluorescent cyanine dye label as the experimental sample, and T-terminated sequencing fragments from M13 labeled with CY5.5 fluorescent cyanine dye label as the reference sequence. Lanes 2, 6, 10 and 14 were loaded with a mixture of C-terminated sequencing fragments from M13 labeled with CY5.0 fluorescent cyanine dye label as the experimental sample and CY5.5-labeled M13 T's as the reference sequence fragments. Lanes 3, 7, 11 and 15 were loaded with a mixture of G-terminated sequencing fragments from M13 labeled with CY5.0 fluorescent cyanine dye label as the experimental sample and CY5.5-labeled M13 T's as the reference sequence. Lanes 4, 8, 12 and 16 were loaded with a mixture of T-terminated sequencing fragments from M13 labeled with CY5.0 fluorescent cyanine dye label as the experimental sample and CY5.5-labeled M13 T's as the reference sample. The reference sequence and the experimental sequence in this example are derived from the same source, and indeed in the case of the T-terminated sequencing fragments are identical to the reference sequence except for the difference in label. However, the good results for alignment and linearization indicate that the reference sequence does not have to be related to the experimental sequence in any way.

The labeled DNA molecules were separated by electrophoresis and detected using a 638 nm laser excitation source which was detected in real time. The data collection was performed on a 2-color DNA Sequencer (Visible Genetics Inc.), to record two channels for each

physical lane, one channel reflecting detection of the CY5.0 label affixed to the experimental sequencing fragments and one channel reflecting detection of the CY5.5 label affixed to the reference fragments. Collected data from the two channels were corrected for overlap in the emission spectra of the two labels and the two resulting data traces were saved as a "data file."

5 Data analysis on the data file was performed using special software in accordance with the protocols of the present invention.

For each reference channel, several peaks (from 3 to 40 in different experiments) were identified having sizes in the range from 40 to about 1200 bases. The base number as assigned to each of these peaks based on knowledge of the sequence of the reference sample, and the position of each peak in the time scale of the experiment was determined. The information about these peaks in the form of a base number and a peak position (or time) was stored in a "peak file."

To align the raw data stored in the data file, the peak data was used to calculate the standard number of bases per unit time as an average over the 16 reference channels. The data was fit to 3<sup>rd</sup> and 5<sup>th</sup> order polynomials expressing the relationship base number and peak position. Using the fitted polynomial, a corrected time scale was created, so that the reference peaks are equally spaced in the corrected time and have the same origin. The number of bases per unit corrected time is constant for all the data in the run. However, the actual time interval between peaks is not generally constant. Thus, the corrected time scale is used to resample the experimental data trace and the associated reference channel. This procedure essentially involves looking at the experimental at the times specified by the corrected time scale, and determining whether or not a peak is present at the correct time.

Figs. 1 and 2 illustrate the application of the invention to the specific sequences described above. Fig. 1 shows the spacing between adjacent bases as a function of base number, for non-aligned (raw) data (closed diamonds), and data aligned and linearized using a 3<sup>rd</sup> order (open triangles) and 5<sup>th</sup> order (open circles) polynomials. . It is clearly seen that the spacing

is changing during the run significantly, but is linearized by fitting with either the 3<sup>rd</sup> or 5<sup>th</sup> order polynomial.

Fig. 2 illustrates the influence of the order of the polynomial used for fitting the raw data of the experimental traces. Increasing the polynomial from 3<sup>rd</sup> order (open triangles) to 4<sup>th</sup> order (closed diamonds) improves linearity noticeably, although the curve still have a nonlinear part in the beginning of the run (up to about 100 bases). The 5<sup>th</sup> order polynomial (open circles) gives the best result, with the maximum deviation from the straight line being less than about 0.5 seconds up to 1300 bases. Such linearity is close to the limit in this particular experiment, because the sampling time was 0.5 seconds. Thus, further increase in the order of the polynomial would only increase computational time, without being likely to provide any significant improvement in linearity.

Figs. 3A and B illustrate improvement in the alignment of the sequencing data (from trace to trace) based on the procedure of the invention. Fig. 3A shows raw data. The difference in run time can reach 500 seconds. Alignment of the raw data, even with a 3<sup>rd</sup>-degree polynomial, improves the data significantly, reducing the difference in run time to a maximum of ~ 90 seconds. (See. Fig. 3B) When a 5<sup>th</sup>-degree polynomial is used, the difference becomes less than 10 seconds.

### Example 2

Raw data traces were generated using M13 T-terminated sequencing fragments in four adjacent lanes of a sequencing gel. As noted in Table 1, the raw, unaligned data traces showed the substantial variability in peak position that can be observed. Application of a 5<sup>th</sup> order polynomial to this data to determine a corrected time scale, and the application of this time scale to the raw data traces, resulted in a substantial improvement in the alignment of the data. This improved alignment allows the calling of bases with greater accuracy over the entire 1300 bases length.

Table 1		
Peak Number	Separation between high and low time peaks, before alignment	Separation between high and low time peaks, after alignment
40	1 min 16 sec	14 sec
140	2 min 5 sec	4 sec
312	9 min 30 sec	6 sec
607	41 min 38 sec	4 sec
970	almost 1.5 hours	24 sec

### Example 3

To understand the significance of the number of peaks incorporated in the peak file for use in generating the polynomial, the data from a T-terminated M13 fragment set was processed using 3, 5, 10, 20 and 40 selected peaks, and the spacing between adjacent peaks at various base positions after alignment was determined. The results are shown in Table 2. As can be seen higher numbers of peaks reduce the extent of variation in peak spacing, although even as few as 3 peaks provides useful results. Comparison of the results from 10, 20 and 40 peaks suggests that an increase beyond 40 would only add to the computational burden without improving the quality of the result.

### Example 4

To evaluate the ability of the linearization and alignment processes of the invention provide a demonstrable improvement in base calling accuracy and read length, M13 sequence was used. CY5.0-labeled A, C, G and T-terminated sequence fragments were used as experimental samples, while M13 T's labeled with CY5.5 were used as the reference sample. Base-calling was performed on the raw data, and on the data after alignment based on 40 peaks of the reference trace.

Table 2

Spacing between adj. Peaks

# OF PEAKS BASE used for Align. #	3pk	5pk	10pk	20pk	40pk
40	10.4	14.6	15.2	14.1	15.7
62	11.1	15.3	15.8	14.6	16.1
95	12.4	16.4	16.7	15.6	17.5
117	13.0	16.7	16.9	15.8	17.4
140	13.7	16.7	16.7	15.6	17.4
194	13.0	16.9	16.7	15.7	17.6
254	16.4	16.9	16.5	15.6	17.6
312	17.3	16.7	16.3	15.4	17.5
331	17.7	16.3	16.0	15.1	17.2
392	18.5	16.3	16.2	15.2	17.3
446	19.5	16.5	16.6	15.6	17.2
519	19.6	16.4	16.7	15.6	17.3
579	19.2	16.4	16.7	15.5	17.3
622	18.7	16.5	16.6	15.6	17.6
701	18.0	16.6	16.5	15.4	17.6
741	17.3	16.7	16.4	15.4	17.7
809	16.4	16.7	16.3	15.3	17.4
882	15.8	17.0	16.6	15.7	17.1
922	15.2	16.8	16.7	15.8	17.0
970	14.1	16.2	16.5	15.5	17.0
1026	13.4	15.7	16.5	15.4	16.9
1047	12.3	15.4	16.5	15.4	16.9
Average	15.7	16.3	16.4	15.4	17.2
SQDEV	7.6	0.3	0.1	0.1	0.2
Stndrd. Dev	2.8	0.6	0.4	0.4	0.5
Max dev	9.2	2.4	1.6	1.7	2.0



The relationship between accuracy and read length for each of these two experiments is shown in Figs. 4 and 5, respectively. As shown in Fig. 4, for a given accuracy (for example 97%), data alignment based on information from a reference channel allows increase in read length for at least 10%, i.e., for another 100 bases to be accurately read.

5 Alternatively, for a given read length (for example 900 bases), it provides improved accuracy (98.5% from 97%). These conclusion are based on results of base calling for lanes that were relatively well-aligned to begin with. For channels which experience a large shift in the raw data, the effect of alignment in accordance with the invention is more pronounced. (Fig. 5). Thus, in this experimental system without alignment it is possible to call only 100 bases with reasonable  
10 accuracy. After alignment, however, up to 1000 bases can be called.

What is Claimed is:

1                   1.       A method for assignment of base numbers to peaks within an experimental  
2 DNA sequencing data trace derived from the separation of experimental DNA sequencing  
3 fragments, comprising the steps of:

4                   (a)       obtaining one or more reference DNA sequencing data traces derived from  
5 the separation of reference DNA sequencing fragments reflecting the position of at least one base  
6 in a reference polynucleotide of known sequence;

7                   (b)       evaluating the reference DNA sequencing data traces to determine a  
8 corrected time scale indicative of migration times at which peaks should occur;

9                   (c)       sampling the experimental DNA sequencing data trace at time points  
10 determined by the corrected time scale, and

11                   (d)       assigning a base number to each peak found in the experimental DNA  
sequencing data trace based upon the corrected time scale.

12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65  
66  
67  
68  
69  
70  
71  
72  
73  
74  
75  
76  
77  
78  
79  
80  
81  
82  
83  
84  
85  
86  
87  
88  
89  
90  
91  
92  
93  
94  
95  
96  
97  
98  
99  
100  
101  
102  
103  
104  
105  
106  
107  
108  
109  
110  
111  
112  
113  
114  
115  
116  
117  
118  
119  
120  
121  
122  
123  
124  
125  
126  
127  
128  
129  
130  
131  
132  
133  
134  
135  
136  
137  
138  
139  
140  
141  
142  
143  
144  
145  
146  
147  
148  
149  
150  
151  
152  
153  
154  
155  
156  
157  
158  
159  
160  
161  
162  
163  
164  
165  
166  
167  
168  
169  
170  
171  
172  
173  
174  
175  
176  
177  
178  
179  
180  
181  
182  
183  
184  
185  
186  
187  
188  
189  
190  
191  
192  
193  
194  
195  
196  
197  
198  
199  
200  
201  
202  
203  
204  
205  
206  
207  
208  
209  
210  
211  
212  
213  
214  
215  
216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230  
231  
232  
233  
234  
235  
236  
237  
238  
239  
240  
241  
242  
243  
244  
245  
246  
247  
248  
249  
250  
251  
252  
253  
254  
255  
256  
257  
258  
259  
260  
261  
262  
263  
264  
265  
266  
267  
268  
269  
270  
271  
272  
273  
274  
275  
276  
277  
278  
279  
280  
281  
282  
283  
284  
285  
286  
287  
288  
289  
290  
291  
292  
293  
294  
295  
296  
297  
298  
299  
300  
301  
302  
303  
304  
305  
306  
307  
308  
309  
310  
311  
312  
313  
314  
315  
316  
317  
318  
319  
320  
321  
322  
323  
324  
325  
326  
327  
328  
329  
330  
331  
332  
333  
334  
335  
336  
337  
338  
339  
340  
341  
342  
343  
344  
345  
346  
347  
348  
349  
350  
351  
352  
353  
354  
355  
356  
357  
358  
359  
360  
361  
362  
363  
364  
365  
366  
367  
368  
369  
370  
371  
372  
373  
374  
375  
376  
377  
378  
379  
380  
381  
382  
383  
384  
385  
386  
387  
388  
389  
390  
391  
392  
393  
394  
395  
396  
397  
398  
399  
400  
401  
402  
403  
404  
405  
406  
407  
408  
409  
410  
411  
412  
413  
414  
415  
416  
417  
418  
419  
420  
421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431  
432  
433  
434  
435  
436  
437  
438  
439  
440  
441  
442  
443  
444  
445  
446  
447  
448  
449  
450  
451  
452  
453  
454  
455  
456  
457  
458  
459  
460  
461  
462  
463  
464  
465  
466  
467  
468  
469  
470  
471  
472  
473  
474  
475  
476  
477  
478  
479  
480  
481  
482  
483  
484  
485  
486  
487  
488  
489  
490  
491  
492  
493  
494  
495  
496  
497  
498  
499  
500  
501  
502  
503  
504  
505  
506  
507  
508  
509  
510  
511  
512  
513  
514  
515  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539  
540  
541  
542  
543  
544  
545  
546  
547  
548  
549  
550  
551  
552  
553  
554  
555  
556  
557  
558  
559  
560  
561  
562  
563  
564  
565  
566  
567  
568  
569  
570  
571  
572  
573  
574  
575  
576  
577  
578  
579  
580  
581  
582  
583  
584  
585  
586  
587  
588  
589  
590  
591  
592  
593  
594  
595  
596  
597  
598  
599  
600  
601  
602  
603  
604  
605  
606  
607  
608  
609  
610  
611  
612  
613  
614  
615  
616  
617  
618  
619  
620  
621  
622  
623  
624  
625  
626  
627  
628  
629  
630  
631  
632  
633  
634  
635  
636  
637  
638  
639  
640  
641  
642  
643  
644  
645  
646  
647  
648  
649  
650  
651  
652  
653  
654  
655  
656  
657  
658  
659  
660  
661  
662  
663  
664  
665  
666  
667  
668  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701  
702  
703  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755  
756  
757  
758  
759  
760  
761  
762  
763  
764  
765  
766  
767  
768  
769  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
780  
781  
782  
783  
784  
785  
786  
787  
788  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809  
810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863  
864  
865  
866  
867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917  
918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971  
972  
973  
974  
975  
976  
977  
978  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000  
1001  
1002  
1003  
1004  
1005  
1006  
1007  
1008  
1009  
1010  
1011  
1012  
1013  
1014  
1015  
1016  
1017  
1018  
1019  
1020  
1021  
1022  
1023  
1024  
1025  
1026  
1027  
1028  
1029  
1030  
1031  
1032  
1033  
1034  
1035  
1036  
1037  
1038  
1039  
1040  
1041  
1042  
1043  
1044  
1045  
1046  
1047  
1048  
1049  
1050  
1051  
1052  
1053  
1054  
1055  
1056  
1057  
1058  
1059  
1060  
1061  
1062  
1063  
1064  
1065  
1066  
1067  
1068  
1069  
1070  
1071  
1072  
1073  
1074  
1075  
1076  
1077  
1078  
1079  
1080  
1081  
1082  
1083  
1084  
1085  
1086  
1087  
1088  
1089  
1090  
1091  
1092  
1093  
1094  
1095  
1096  
1097  
1098  
1099  
1100  
1101  
1102  
1103  
1104  
1105  
1106  
1107  
1108  
1109  
1110  
1111  
1112  
1113  
1114  
1115  
1116  
1117  
1118  
1119  
1120  
1121  
1122  
1123  
1124  
1125  
1126  
1127  
1128  
1129  
1130  
1131  
1132  
1133  
1134  
1135  
1136  
1137  
1138  
1139  
1140  
1141  
1142  
1143  
1144  
1145  
1146  
1147  
1148  
1149  
1150  
1151  
1152  
1153  
1154  
1155  
1156  
1157  
1158  
1159  
1160  
1161  
1162  
1163  
1164  
1165  
1166  
1167  
1168  
1169  
1170  
1171  
1172  
1173  
1174  
1175  
1176  
1177  
1178  
1179  
1180  
1181  
1182  
1183  
1184  
1185  
1186  
1187  
1188  
1189  
1190  
1191  
1192  
1193  
1194  
1195  
1196  
1197  
1198  
1199  
1200  
1201  
1202  
1203  
1204  
1205  
1206  
1207  
1208  
1209  
1210  
1211  
1212  
1213  
1214  
1215  
1216  
1217  
1218  
1219  
1220  
1221  
1222  
1223  
1224  
1225  
1226  
1227  
1228  
1229  
1230  
1231  
1232  
1233  
1234  
1235  
1236  
1237  
1238  
1239  
1240  
1241  
1242  
1243  
1244  
1245  
1246  
1247  
1248  
1249  
1250  
1251  
1252  
1253  
1254  
1255  
1256  
1257  
1258  
1259  
1260  
1261  
1262  
1263  
1264  
1265  
1266  
1267  
1268  
1269  
1270  
1271  
1272  
1273  
1274  
1275  
1276  
1277  
1278  
1279  
1280  
1281  
1282  
1283  
1284  
1285  
1286  
1287  
1288  
1289  
1290  
1291  
1292  
1293  
1294  
1295  
1296  
1297  
1298  
1299  
1300  
1301  
1302  
1303  
1304  
1305  
1306  
1307  
1308  
1309  
1310  
1311  
1312  
1313  
1314  
1315  
1316  
1317  
1318  
1319  
1320  
1321  
1322  
1323  
1324  
1325  
1326  
1327  
1328  
1329  
1330  
1331  
1332  
1333  
1334  
1335  
1336  
1337  
1338  
1339  
1340  
1341  
1342  
1343  
1344  
1345  
1346  
1347  
1348  
1349  
1350  
1351  
1352  
1353  
1354  
1355  
1356  
1357  
1358  
1359  
1360  
1361  
1362  
1363  
1364  
1365  
1366  
1367  
1368  
1369  
1370  
1371  
1372  
1373  
1374  
1375  
1376  
1377  
1378  
1379  
1380  
1381  
1382  
1383  
1384  
1385  
1386  
1387  
1388  
1389  
1390  
1391  
1392  
1393  
1394  
1395  
1396  
1397  
1398  
1399  
1400  
1401  
1402  
1403  
1404  
1405  
1406  
1407  
1408  
1409  
1410  
1411  
1412  
1413  
1414  
1415  
1416  
1417  
1418  
1419  
1420  
1421  
1422  
1423  
1424  
1425  
1426  
1427  
1428  
1429  
1430  
1431  
1432  
1433  
1434  
1435  
1436  
1437  
1438  
1439  
1440  
1441  
1442  
1443  
1444  
1445  
1446  
1447  
1448  
1449  
1450  
1451  
1452  
1453  
1454  
1455  
1456  
1457  
1458  
1459  
1460  
1461  
1462  
1463  
1464  
1465  
1466  
1467  
1468  
1469  
1470  
1471  
1472  
1473  
1474  
1475  
1476  
1477  
1478  
1479  
1480  
1481  
1482  
1483  
1484  
1485  
1486  
1487  
1488  
1489  
1490  
1491  
1492  
1493  
1494  
1495  
1496  
1497  
1498  
1499  
1500  
1501  
1502  
1503  
1504  
1505  
1506  
1507  
1508  
1509  
1510  
1511  
1512  
1513  
1514  
1515  
1516  
1517  
1518  
1519  
1520  
1521  
1522  
1523  
1524  
1525  
1526  
1527  
1528  
1529  
1530  
1531  
1532  
1533  
1534  
1535  
1536  
1537  
1538  
1539  
1540  
1541  
1542  
1543  
1544  
1545  
1546  
1547  
1548  
1549  
1550  
1551  
1552  
1553  
1554  
1555  
1556  
1557  
1558  
1559  
1560  
1561  
1562  
1563  
1564  
1565  
1566  
1567  
1568  
1569  
1570  
1571  
1572  
1573  
1574  
1575  
1576  
1577  
1578  
1579  
1580  
1581  
1582  
1583  
1584  
1585  
1586  
1587  
1588  
1589  
1590  
1591  
1592  
1593  
1594  
1595  
1596  
1597  
1598  
1599  
1600  
1601  
1602  
1603  
1604  
1605  
1606  
1607  
1608  
1609  
1610  
1611  
1612  
1613  
1614  
1615  
1616  
1617  
1618  
1619  
1620  
1621  
1622  
1623  
1624  
1625  
1626  
1627  
1628  
1629  
1630  
1631  
1632  
1633  
1634  
1635  
1636  
1637  
1638  
1639  
1640  
1641  
1642  
1643  
1644  
1645  
1646  
1647  
1648  
1649  
1650  
1651  
1652  
1653  
1654  
1655  
1656  
1657  
1658  
1659  
1660  
1661  
1662  
1663  
1664  
1665  
1666  
1667  
1668  
1669  
1670  
1671  
1672  
1673  
1674  
1675  
1676  
1677  
1678  
1679  
1680  
1681  
1682  
1683  
1684  
1685  
1686  
1687  
1688  
1689  
1690  
1691  
1692  
1693  
1694  
1695  
1696  
1697  
1698  
1699  
1700  
1701  
1702  
1703  
1704  
1705  
1706  
1707  
1708  
1709  
1710  
1711  
1712  
1713  
1714  
1715  
1716  
1717  
1718  
1719  
1720  
1721  
1722  
1723  
1724  
1725  
1726  
1727  
1728  
1729  
1730  
1731  
1732  
1733  
1734  
1735  
1736  
1737  
1738  
1739  
1740  
1741  
1742  
1743  
1744  
1745  
1746  
1747  
1748  
1749  
1750  
1751  
1752  
1753  
1754  
1755  
1756  
1757  
1758  
1759  
1760  
1761  
1762  
1763  
1764  
1765  
1766  
1767  
1768  
1769  
1770  
1771  
1772  
1773  
1774  
1775  
1776  
1777  
1778  
1779  
1780  
1781  
1782  
1783  
1784  
1785  
1786  
1787  
1788  
1789  
1790  
1791  
1792  
1793  
1794  
1795  
1796  
1797  
1798  
1799  
1800  
1801  
1802  
1803  
1804  
1805  
1806  
1807  
1808  
1809  
1810  
1811  
1812  
1813  
1814  
1815  
1816  
1817  
1818  
1819  
1820  
1821  
1822  
1823  
1824  
1825  
1826  
1827  
1828  
1829  
1830  
1831  
1832  
1833  
1834  
1835  
1836  
1837  
1838  
1839  
1840  
1841  
1842  
1843  
1844  
1845  
1846  
1847  
1848  
1849  
1850  
1851  
1852  
1853  
1854  
1855  
1856  
1857  
1858  
1859  
1860  
1861  
1862  
1863  
1864  
1865  
1866  
1867  
1868  
1869  
1870  
1871  
1872  
1873  
1874  
1875  
1876  
1877  
1878  
1879  
1880  
1881  
1882  
1883  
1884  
1885  
1886  
1887  
1888  
1889  
1890  
1891  
1892  
1893  
1894  
1895  
1896  
1897  
1898  
1899  
1900  
1901  
1902  
1903  
1904  
1905  
1906  
1907  
1908  
1909  
1910  
1911  
1912  
1913  
1914  
1915  
1916  
1917  
1918  
1919  
1920  
1921  
1922  
1923  
1924  
1925  
1926  
1927  
1928  
1929  
1930  
1931  
1932  
1933  
1934  
1935  
1936  
1937  
1938  
1939  
1940  
1941  
1942  
1943  
1944  
1945  
1946  
1947  
1948  
1949  
1950  
1951  
1952  
1953  
1954  
1955  
1956  
1957  
1958  
1959  
1960  
1961  
1962  
1963  
1964  
1965  
1966  
1967  
1968  
1969  
1970  
1971  
1972  
1973  
1974  
1975  
1976  
1977  
1978  
1979  
1980  
1981  
1982  
1983  
1984  
1985  
1986  
1987  
1988  
1989  
1990  
1991  
1992  
1993  
1994  
1995  
1996  
1997  
1998  
1999  
2000  
2001  
2002  
2003  
2004  
2005  
2006  
2007  
2008  
2009  
2010  
2011  
2012  
2013  
2014  
2015  
2016  
2017  
2018  
2019  
2020  
2021  
2022  
2023  
2024  
2025  
2026  
2027  
2028  
2029  
2030  
2031  
2032  
2033  
2034  
2035  
2036  
2037  
2038  
2039  
2040  
2041  
2042  
2043  
2044  
2045  
2046  
2047  
2048  
2049  
2050  
2051  
2052  
2053  
2054  
2055  
2056  
2057  
2058  
2059  
2060  
2061  
2062  
2063  
2064  
2065  
2066  
2067  
2068  
2069  
2070  
2071  
2072  
2073  
2074  
2075  
2076  
2077  
2078  
2079  
2080  
2081  
2082  
2083  
2084  
2085  
2086  
2087  
2088  
2089  
2090  
2091  
2092  
2093  
2094  
2095  
2096  
2097  
2098  
2099  
2100  
2101  
2102  
2103  
2104  
2105  
2106  
2107  
2108  
2109  
2110  
2111  
2112  
2113  
2114  
2115  
2116  
2117  
2118  
2119  
2120  
2121  
2122  
2123  
2124  
2125  
2126  
2127  
2128  
2129  
2130  
2131  
2132  
2133  
2134  
2135  
2136  
2137  
2138  
2139  
2140  
2141  
2142  
2143  
2144  
2145  
2146  
2147  
2148  
2149  
2150  
2151  
2152  
2153  
2154  
2155  
2156  
2157  
2158  
2159  
2160  
2161  
2162  
2163  
2164  
2165  
2166  
2167  
2168  
2169  
2170  
2171  
2172  
2173  
2174  
21

1                   3.       The method of claim 1, wherein the experimental DNA sequencing data  
2 trace and a first reference DNA sequencing data trace are derived from analysis of sequencing  
3 fragments in a common lane of a sequencing gel.

1                   4.       The method of claim 1, wherein a plurality of reference DNA sequencing  
2 data traces are obtained, each derived from the separation of the same set of reference DNA  
3 sequencing fragments.

1                   5.       The method of claim 1, wherein the polynomial is a third or higher order  
2 polynomial.

1                   6.       The method of claim 1, wherein a defined number of bands are selected  
2 for evaluation from each of the reference DNA sequencing data traces.

1                   7.       The method of claim 6, wherein the defined number of bands selected is  
2 from 3 to 40.

1                   8.       The method of claim 6, wherein the defined number of bands is at least  
2 equal to the order of the polynomial, plus 1.

1                   9.       The method of claim 1, wherein base numbers are assigned to peaks  
2 within a plurality of experimental DNA sequencing data traces derived from the separation of  
3 experimental DNA sequencing fragments indicative of the positions of a plurality of types of  
4 bases.

1           10.    The method of claim 9, wherein base numbers are assigned to peaks  
2    within four experimental DNA sequencing data traces derived from the separation of  
3    experimental DNA sequencing fragments indicative of the positions of four types of bases.

1           11.    A method for evaluating the sequence of a target polynucleotide,  
2    comprising the steps of:

3           (a)    obtaining one or more experimental DNA sequencing data traces derived  
4    from the separation of experimental DNA sequencing fragments reflecting the position of at least  
5    one base in the target polynucleotide and one or more reference DNA sequencing data traces  
6    derived from the separation of reference DNA sequencing fragments reflecting the position of at  
7    least one base in a reference polynucleotide of known sequence;

8           (b)    evaluating the reference DNA sequencing data traces to determine a  
9    corrected time scale indicative of migration times at which peaks should occur;

10          (c)    sampling the experimental DNA sequencing data traces at time points  
11   determined by the corrected time scale, and

12          (d)    assigning a base number to each peak found in the experimental DNA  
13   sequencing data traces based upon the corrected time scale, thereby obtaining information about  
14   the sequence of the target polynucleotide.

1           12.    The method of claim 11, wherein the step of evaluating the reference DNA  
2    sequence data traces includes the steps of:

3           (i)    identifying a plurality of peaks in the reference DNA sequencing data  
4    traces, and creating a data table containing the number of each peak based on the known  
5    sequence of the polynucleotide, and the position of each peak in the reference DNA sequencing  
6    data trace;

7           (ii)   identifying a set of coefficients for a polynomial effective to substantially  
8   linearize a plot of peak number versus separation between adjacent peaks; and

9 (iii) creating from the coefficients and the polynomial a corrected time scale  
10 which reflects the positions at which a peak should occur at any given point in a sequencing data  
11 trace.

1 13. The method of claim 11, wherein the reference DNA sequencing traces  
2 and the experimental DNA sequencing data trace are derived from analysis of sequencing  
3 fragments in a common sequencing gel.

1 14. The method of claim 13, wherein the experimental DNA sequencing data  
2 trace and a first reference DNA sequencing data trace are derived from analysis of sequencing  
3 fragments in a common lane of the common sequencing gel.

1 15. The method of claim 11, wherein a plurality of reference DNA sequencing  
2 data traces are obtained, each derived from the separation of the same set of reference DNA  
3 sequencing fragments.

1 16. The method of claim 11, wherein the polynomial is a third or higher order  
2 polynomial.

1 17. The method of claim 11, wherein a defined number of bands are selected  
2 for evaluation from each of the reference DNA sequencing data traces.

1 18. The method of claim 17, wherein the defined number of bands selected is  
2 from 3 to 40.

1 19. The method of claim 17, wherein the defined number of bands is at least  
2 equal to the order of the polynomial, plus 1.

1                   20.     The method of claim 11, wherein base numbers are assigned to peaks  
2     within a plurality of experimental DNA sequencing data traces derived from the separation of  
3     experimental DNA sequencing fragments indicative of the positions of a plurality of types of  
4     bases.

1                   21.     An apparatus for evaluating the sequence of a target polynucleotide,  
2     comprising:

3                   (a)     an input for receiving information about one or more experimental DNA  
4     sequencing data traces derived from the separation of experimental DNA sequencing fragments  
5     reflecting the position of at least one base in the target polynucleotide and one or more reference  
6     DNA sequencing data traces derived from the separation of reference DNA sequencing  
7     fragments reflecting the position of at least one base in a reference polynucleotide of known  
8     sequence;

9                   (b)     a processor, operatively programmed to evaluate the reference DNA  
10     sequencing data traces to determine a corrected time scale indicative of migration times at which  
11     peaks should occur;

12                  (c)     a processor, operatively programed to sample the experimental DNA  
13     sequencing data traces at time points determined by the corrected time scale;

14                  (d)     a processor, operatively programmed to assign a base number to each peak  
15     found in the experimental DNA sequencing data traces based upon the corrected time scale,  
16     thereby obtaining information about the sequence of the target polynucleotide; and

17                  (e)     an output for communicating the information about the sequence of the  
18     target polynucleotide.

1                   22.     The apparatus of claim 21, wherein the processor programmed to evaluate  
2     the reference DNA sequence data traces is programmed to perform the steps of:

3                   (i)     identifying a plurality of peaks in the reference DNA sequencing data  
4     traces, and creating a data table containing the number of each peak based on the known  
5     sequence of the polynucleotide, and the position of each peak in the reference DNA sequencing  
6     data trace;

7                   (ii)    identifying a set of coefficients for a polynomial effective to substantially  
8     linearize a plot of peak number versus separation between adjacent peaks; and

9                   (iii)   creating from the coefficients and the polynomial a corrected time scale  
10    which reflects the positions at which a peak should occur at any given point in a sequencing data  
11    trace.

ABSTRACT OF THE DISCLOSURE

1 In order to align DNA sequence data traces, an experimental data trace  
2 representing the positions of a first species of base within a target polynucleotide and a reference  
3 data trace representing the positions of a second species of base (which may be the same as or  
4 different from the first species) within a reference polynucleotide are obtained by separating  
5 appropriate sequencing fragments generated from the target and reference polynucleotides on an  
6 electrophoresis gel. For each reference data trace, a plurality of peaks corresponding to  
7 fragments having a size in the range of 40 to 1200 bases are selected. A base number is assigned  
8 to each of the selected peaks in the reference data trace, and a numerical "peak file" is created  
9 with information about the peak number and migration time (or distance). This peak file is  
10 analyzed to determine a set of polynomial coefficients which will allow substantial linearization  
11 of a plot of peak number versus separation between adjacent peaks and alignment of the traces  
12 with respect to each other. These coefficients are used to create a corrected time scale identifying  
13 where peaks should be located on a given experimental gel. This corrected time scale is used to  
14 guide the sampling of the experimental data, and for assignment of peaks within the data.



Spacing Between Adjacent Peaks as a Function of Base Pair Number  
(M13, T's T=6%, 60C, Long Gel, V=1500V)

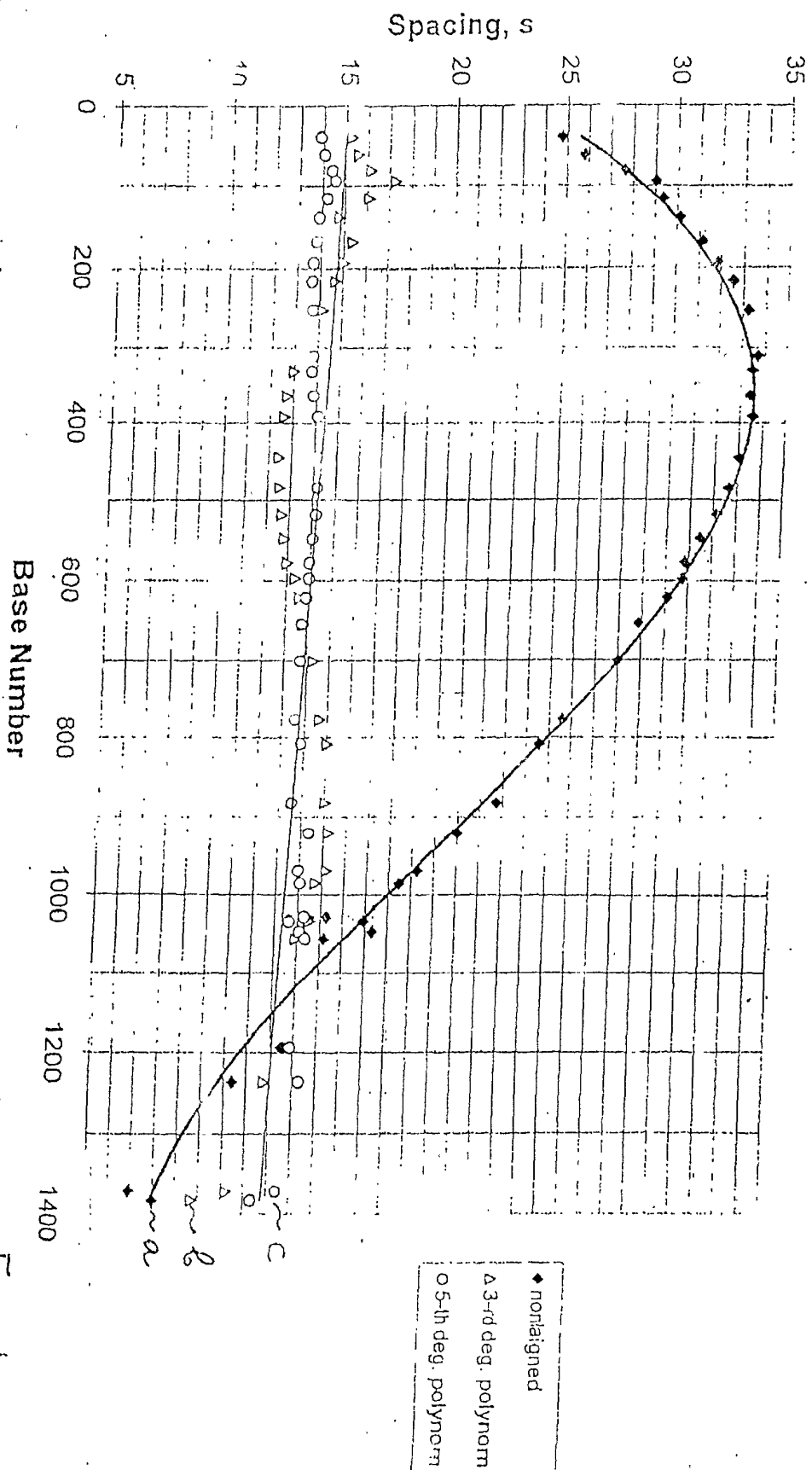


Fig. 1

0053611-032400

Spacing Between Adjacent Peaks as a Function of Base Pair Number  
 (M13, T's T=6%, 60C, Long Gel,  $V = 1500V$ )

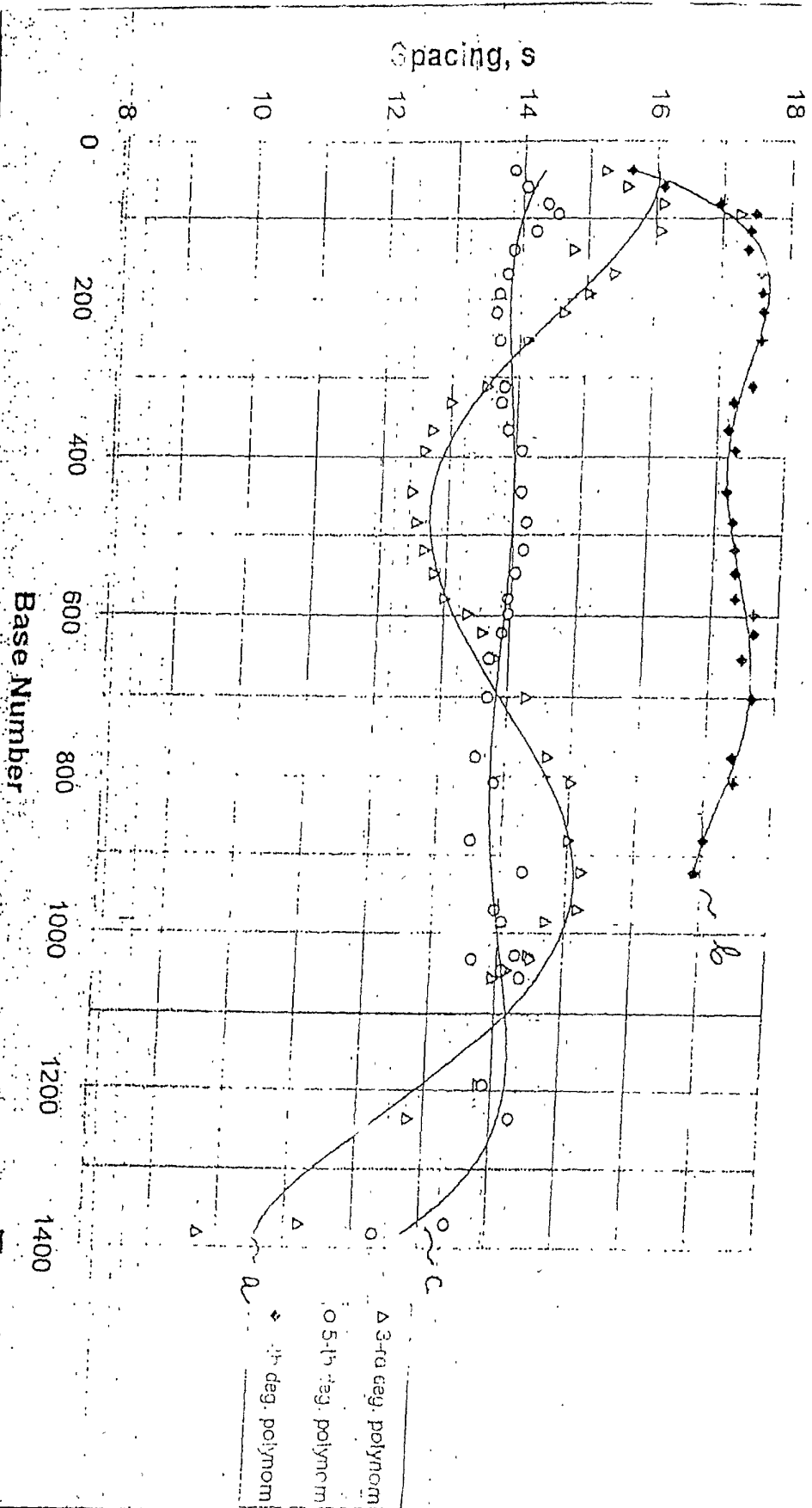


Fig. 2

09536444.032400



Accuracy of Base Calling for the Raw Data and Aligned Data (seq3)  
(data aligned to the reference channel: M13T's 40 peaks)

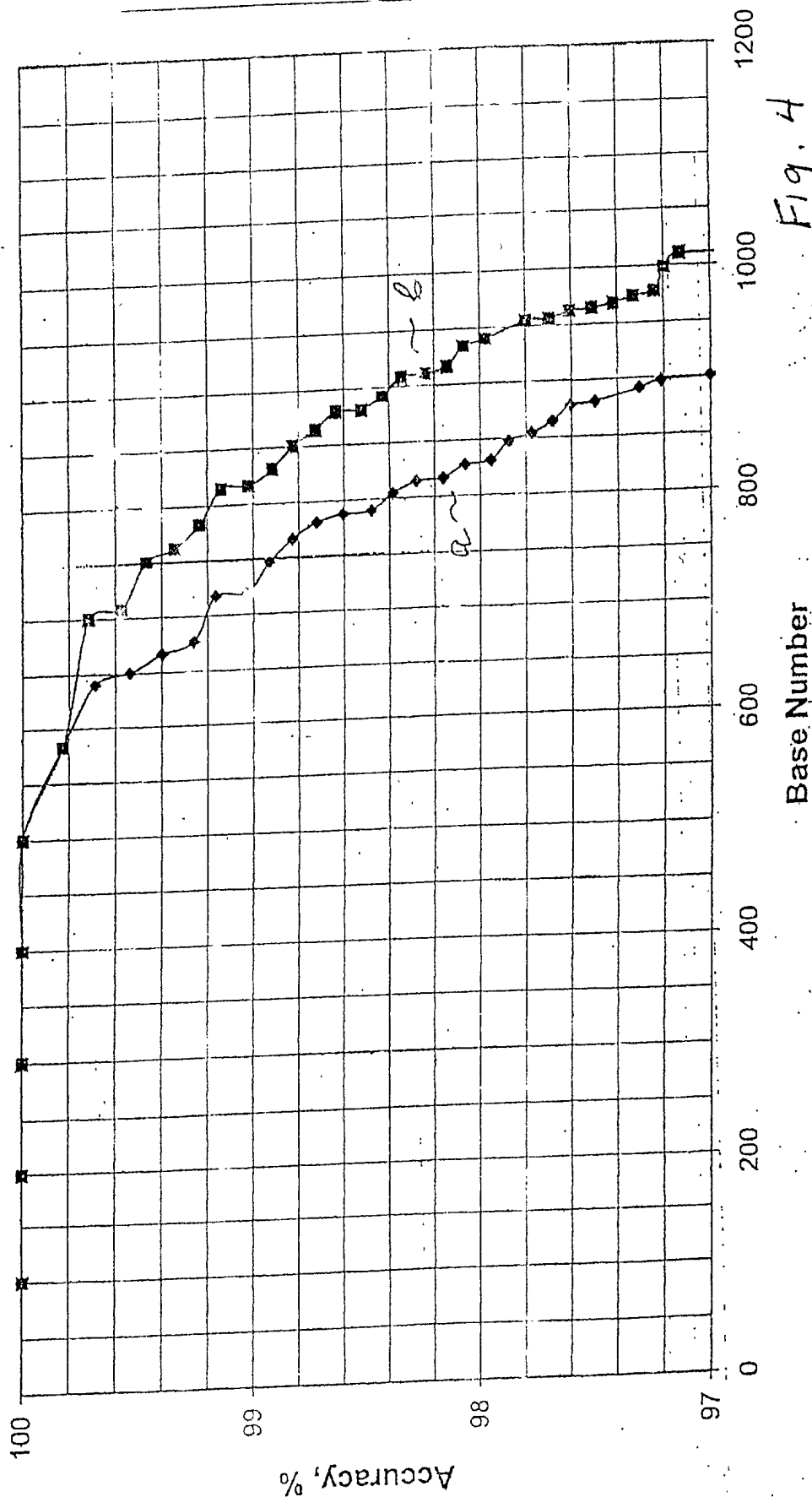


Fig. 4

004220 TT 92550

Accuracy of Base Calling for the Raw Data and Aligned Data (seq1)  
(data aligned to the reference channel: M13T's 40 peaks)

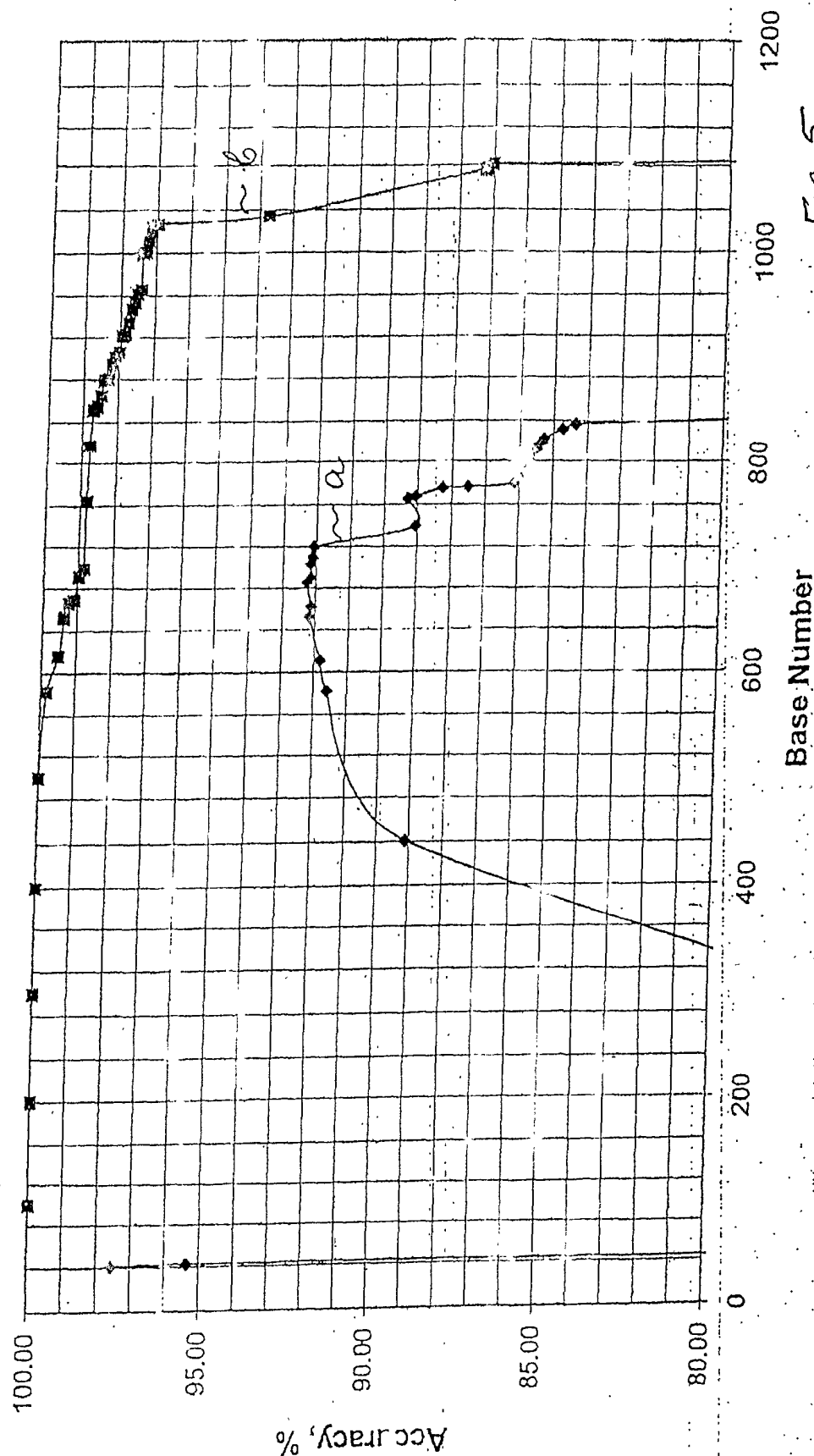


Fig. 5

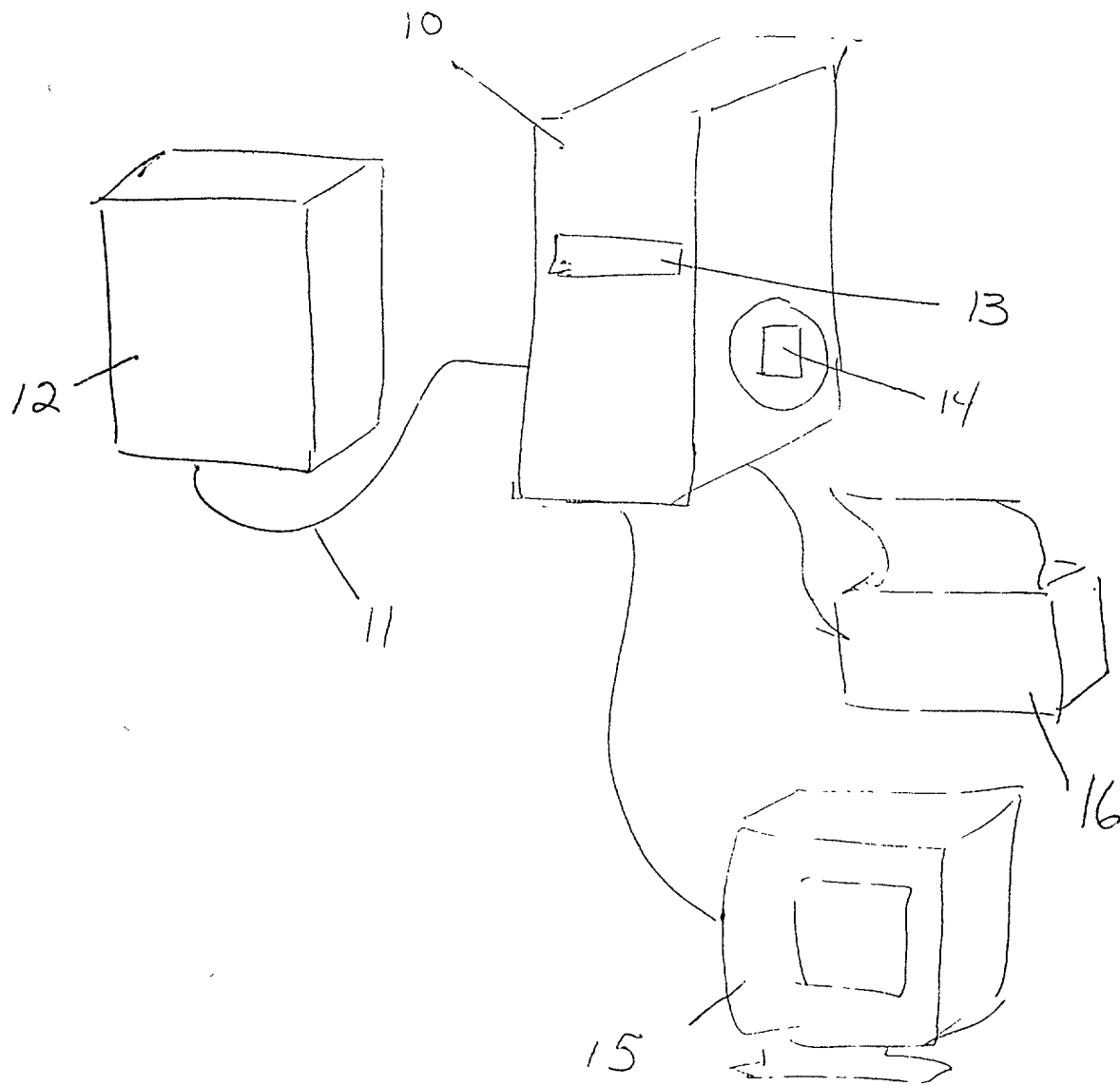


Fig. 6